Volume 20 (2), pp. 161–178
http://www.orssa.org.za

ORiON
ISSN 0529-191-X

Statistical properties of indicators of first-year performance at university

NJ Le Roux* A Bothma† HL Botha‡

Received: 4 August 2004; Revised: 22 October 2004; Accepted: 25 October 2004

Abstract

Appraisal of admission procedures is a matter of urgency for South African universities, as well as for schools producing the prospective students. In this article the focus is on how various statistical procedures can be used to assess admission measures. Properties of the statistical distributions related to school results, access test results and first-year university performance are vital for decision-makers in schools preparing the prospective students and for those who wish to refine university admission procedures. These properties are scrutinised for the 1999, 2000 and 2001 intake groups required to write access tests before being admitted to Stellenbosch University. Using kernel density estimates the univariate distributions of all variables concerned are described in detail. Bagplots are proposed for visual displays of important features like location, spread, correlation, skewness, outliers and tails of bivariate distributions composed of university average performance and a school result or access test variable. Evidence is provided that certain access tests (Mathematics, Science and Numeracy Skills) have statistical distributions similar to that of average first-year university performance, but that average school marks could not be trusted to discriminate between potentially successful and unsuccessful university students.

Key words: Access tests, admission procedures, average first-year university performance, bagplots, bivariate distribution, depth median, kernel density estimates, school marks

1 Introduction

The relationship between school results, first-year university performance and the results of other assessment instruments has been of international interest for many years (think, for example, of the SAT debate ([7] in the USA). The innovative use of statistics to assist in a better understanding of these complex relationships is at the core of this article which is a follow-up of Bothma *et al.* (2004). Bothma *et al.* relate first-year performance at Stellenbosch University (SU) to access test results as well as school results. The authors analysed

*Corresponding author: Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, Stellenbosch, South Africa, email: njlr@sun.ac.za

†Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland, 7602, Stellenbosch, South Africa

‡Academic Support, University of Stellenbosch, Private Bag X1, Matieland, 7602, Stellenbosch, South Africa

first-year performance of the 1999, 2000 and 2001 intake groups, consisting of those applicants who were required to write an access Test Battery prior to being admitted to the University. Three different Test Batteries were administered by SU according to three groups of faculties: students eventually admitted to the Health Sciences, Natural Sciences, Agricultural and Forestry Sciences or Engineering Faculties wrote Test Battery 1; students admitted to the Economic and Management Sciences Faculty wrote Test Battery 2, while those admitted to the Arts, Theology, Law or Education Faculties wrote Test Battery 3. The three Test Batteries were composed of the following individual tests:

Test Battery 1:	Mathematics (Maths); Physical Science (Science); Language (Lang)
Test Battery 2:	Mathematics (Maths); Numeracy Skills (Numer); Language (Lang)
Test Battery 3:	Academic Language Proficiency in Afrikaans (Afr); Academic Language Proficiency in English (Eng); Thinking Skills (Think).

In addition to the access test results the following school results were also considered: the Grade 12 Mathematics mark (Maths.12), the Grade 12 Afrikaans mark (Afr.12) and the Grade 12 English mark (Eng.12) as well as the Grade 11 final average mark (Ave.11) and the Grade 12 final average mark (Ave.12). Bothma *et al.* (2004) provide detailed descriptions of the data sets compiled, as well as the calculation of the first-year weighted university mark (FYWUM).

Although all correlations with FYWUM were relatively small, Bothma *et al.* (2004) demonstrated that FYWUM was more strongly correlated with school result variables than with access test variables. However, boxplots reveal that school results on average lead to unrealistically high expectations of performance at university. In fact, differences as large as 33.92 percentage points were recorded between average FYWUM and Ave.12 marks. Access tests, on the other hand, create a much more realistic idea of a prospective student's performance in the first year at university.

The focus of this follow-up article is on how various statistical procedures may be used to assess university admission measures. The statistical distributions of the different variables considered by Bothma *et al.* (2004) are analysed in detail by means of density estimates of all univariate distributions, as well as sophisticated displays of selected bivariate distributions. The importance of using these density estimates and graphical displays is clearly demonstrated by their ability to uncover several properties and relationships associated with the variables under consideration. Moreover, this article shows how to obtain more accurate probability estimates. It is argued that the density estimates and bivariate graphical displays provide very important additional information for decisions on preparation for and admission procedures at university. Thus the statistical methodologies employed in this research are of particular importance to all researchers in the field of education because they provide insights not apparent when conventional methods are used.

2 Density estimation

The practice of representing univariate distributions of variables in histograms has two major shortcomings: the number of bins as well as the starting point of the grid of bins may have a surprisingly significant effect on the form of the histogram [8, 10]. Various

density estimates addressing these problems have been proposed in the statistical literature [8, 10]. In this article the univariate distributions of variables are estimated by kernel density estimates [10] based on a Gaussian kernel. The bandwidth is chosen to compromise between smoothing sufficiently to remove insignificant bumps and not smoothing so much as to eliminate real peaks.

Consider estimating the density of the 1999 Test Battery 1 variable FYWUM. The probability histogram of this variable is given in Figure 1. It is common practice to fit a normal density with the sample mean and sample variance as parameters to such a histogram. Superimposed on the probability histogram in Figure 1 is this density estimate (solid line). The kernel density estimate is added as a broken line in the figure. This figure illustrates the advantages of the kernel density estimate: it preserves the skewness and bimodality of FYWUM, while the normal estimate forces symmetry and unimodality.

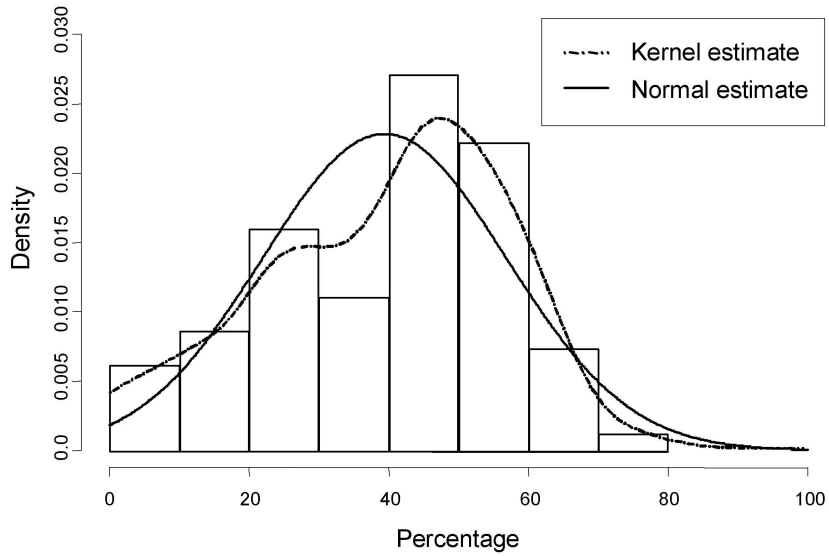

Figure 1: *Histogram with density estimates of the Test Battery 1 variable FYWUM for the 1999 intake group.*

One of the reasons for using normal density estimates is the ease of calculating probabilities. However, it is clear from the graph in of Figure 1 that if the normal density estimate is used for predicting values here, this could result in certain values being overestimated and others being underestimated. Since the kernel density estimate preserves the bimodality as well as the skewness of this data, it should also lead to more accurate predictions in such cases than the normal density estimate could. These predictions can easily be made by means of numerical integration procedures. Therefore, the underlying densities of the access test and school result variables as well as FYWUM are analysed here and probabilities are calculated by means of a kernel density function. This allows conclusions to be drawn without the need for rather restrictive normal assumptions.

2.1 Density estimates of variables associated with Test Battery 1

Figure 2 consists of density estimates of the three access tests variables of Test Battery 1 together with FYWUM for the 1999, 2000 and 2001 intake groups, respectively.

Figure 2 reveals that the modes of the three access tests occur in the same order for each of the three intake groups: Maths has modes that are less than 40%, followed by Science, while the modes of Lang remain larger than 65%. Calculation of the access test means reveals that they can be ordered as the modes, but the mean and mode of a variable may differ extensively. The mean of FYWUM in 2001 is 50%, but its density estimate has a mode at about 65%, with a heavy tail towards the left. This stresses the importance of non-parametric density estimation: several properties of a variable's distribution are revealed without specifying restrictive parametric conditions like unimodality and symmetry.

It is striking that the density estimate of Maths is clearly skewed to the right and that of Lang skewed to the left. Furthermore, the peaks (*i.e.* the observation with highest density) for Maths and Lang remain at approximately 30% and 70%, respectively. Note that the density estimates of Maths have two modes in 2000 and 2001, respectively, and that the second (lower) mode occurs close to the score of 60% in both cases. This suggests that the density of Maths could be a mixture of two or more densities: the first having a mean of approximately 40% and the second having a mean of about 60%. Thus two groups of students are distinguished: a large group with weak mathematical skills and a smaller group with stronger mathematical skills. This could be explained by the fact that only some of the students who wrote the Test Battery 1 access tests will follow a course that requires strong mathematical skills. The usual normal density estimate is unable to extract this information from the data.

The density estimates of Science have a main peak of approximately 40% both in 1999 and 2000, and a secondary peak at approximately 60% in 2000. Note that the density estimate of Science behaves substantially differently in 2001: its form approaches symmetry, with a main peak of about 50% and two smaller humps at about 40% and 70%, respectively.

Since the threshold for exemption of writing the Test Battery 1 access tests was raised for the 2001 intake group, it was to be expected that this would lead to an improvement of the access test results. This is reflected most clearly by changes in the density estimates. The following changes in FYWUM are observed: in 1999 the main peak was about 47%, in 2000 it dropped to approximately 37% and in 2001 the main peak corresponded to a score of 65%. Note that FYWUM is skewed to the left only in 2001, but skewed to the right in 1999 and 2000. The density estimates of Maths and Science also show definite changes in 2001: their variation increased substantially in 2001, and the distribution of Science has a main peak in excess of 50% in 2001, while being skewed to the right in the previous two years.

The kernel density estimates associated with the Grade 11 and Grade 12 final marks and FYWUM of the Test Battery 1 for the 1999, 2000 and 2001 intake groups appear in Figure 3. Note that two different calibrations are used to represent the density (y) axis in Figures 2 and 3 respectively. However, a fixed calibration is used within each figure to enable comparisons between corresponding densities of the three intake groups.

In Figure 3 it is striking that almost all students required to write access tests in the

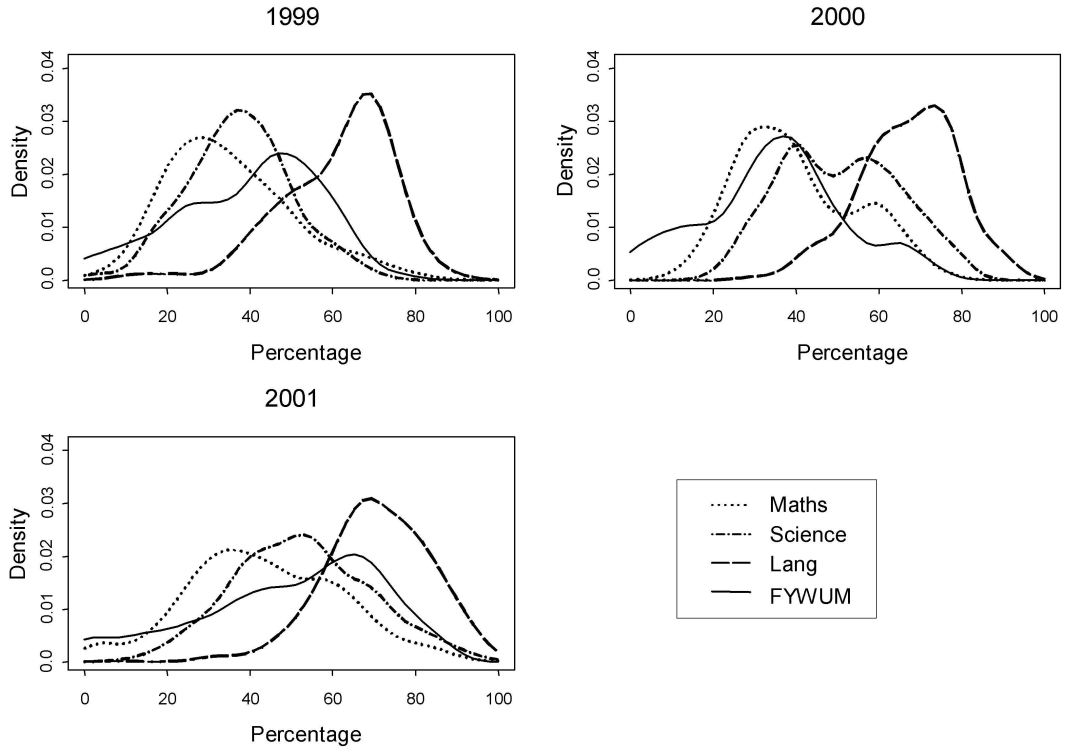

Figure 2: *Kernel density estimates of the densities associated with the access test variables and the FYWUM of the Test Battery 1 data sets for 1999, 2000 and 2001.*

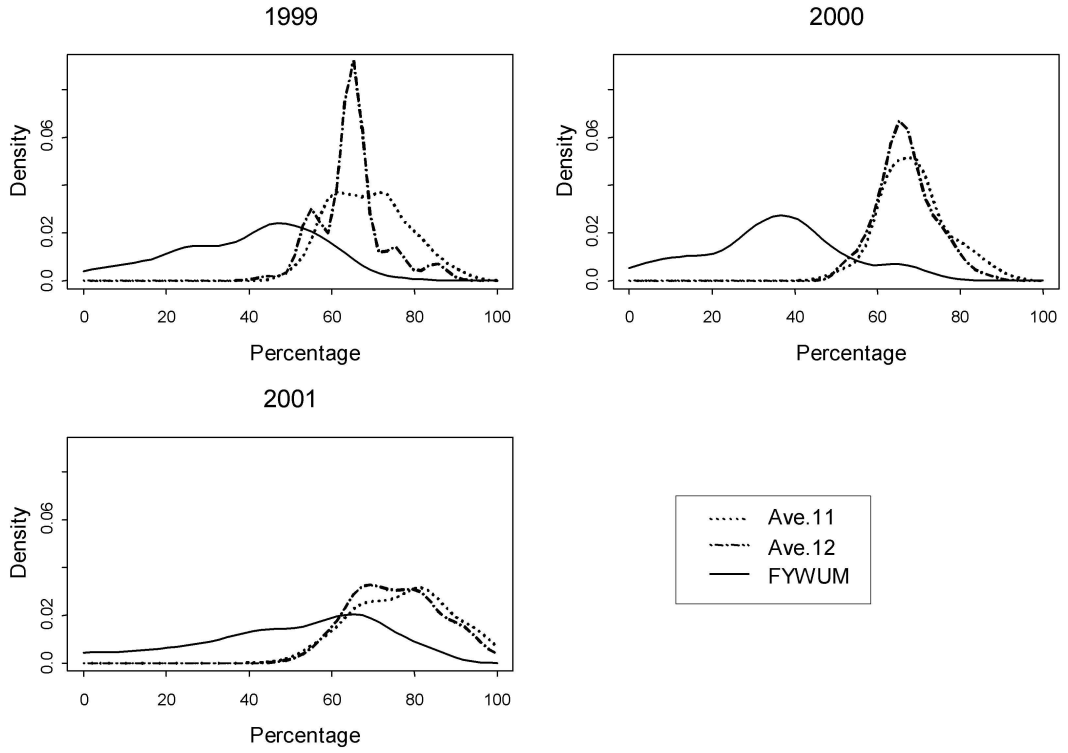

Figure 3: *Kernel density estimates of the densities associated with the Grade 11 and Grade 12 mark variables and the FYWUM of the Test Battery 1 data sets for 1999, 2000 and 2001.*

1999 and 2000 intake groups had Grade 11 and 12 marks in excess of 60%, but more than half of these students achieved a FYWUM of less than 50%. The main peak of Ave.12 is approximately 65% for both 1999 and 2000. The 1999 Ave.11 density estimate has two modes, indicating clearly that the largest number of students in this group has a Grade 11 average mark between 60% and 75%, while in 2000 the main peak occurs at 65%.

Furthermore, it is clear that only in 2001 did the majority of students achieve more than 50% in Ave.11, Ave.12 as well as FYWUM. The main peaks of FYWUM, Ave.11 and Ave.12 correspond to 65%, 70% and 82%, respectively, while the corresponding means are 50%, 77% and 76%. Note that the density estimates of Ave.11 and Ave.12 are bimodal, with the modes towards the right-hand side representing the effect of the stricter criterion of at least 85% Grade 11 final average mark or Grade 12 final average examination mark introduced for the 2001 Health Sciences students before being exempted from writing the access tests. Bothma *et al.* (2004) call this effect the *effect of the Health Sciences 2001 students*. The results of the 2001 intake are considerably different from the results of the previous two years, and it can be concluded that the increase in average Grade 11 and Grade 12 marks of students writing the access tests led to an increase in the average FYWUM.

In view of the Test Battery 1 results, two important aspects need further investigation in the Test Battery 2 and Test Battery 3 data sets: does the distribution of FYWUM remain skewed to the right and what distributions have the closest resemblance with FYWUM?

2.2 Density estimates of variables associated with Test Battery 2

The estimated densities of the access test and school result variables of the Test Battery 2 data sets are compared to those of FYWUM in Figures 4 and 5. It should be kept in mind that the prospective students who wrote the Test Battery 2 access tests had Grade 11 and Grade 12 marks of approximately between 50% and 80%.

It is seen in Figure 4 that the modes of Maths and Numer are less than 40% for the three intake groups, while the modes of Lang exceed 60%. It is noteworthy, again, that the densities reveal far more properties of the underlying distributions of the access test variables than reporting only a location estimate.

A prominent tendency occurs over the three years. More than half of the students have Maths and FYWUM scores of less than 40% as well as Numer scores of less than 50%, but the majority of the students have a Lang score of more than 60%. Note that the density estimates of Maths, Numer and FYWUM show a great deal of similarity, concerning their modes as well as skewness. Since only a small percentage of students had a FYWUM or Maths score higher than 50%, it could be reasoned that if a prospective student achieved a score less than 50% for Maths, the chances of achieving a FYWUM mark in excess of 50% are remote.

The humps to the right of the main peaks of the distributions of Maths, Numer and FYWUM are also noteworthy: it seems that the distributions of these variables are in fact mixtures of two or more distributions. Assuming that each of the distributions of Maths, Numer and FYWUM is a mixture of two normal distributions, it can be said that the normal density with a lower mean has greater weight than the normal density

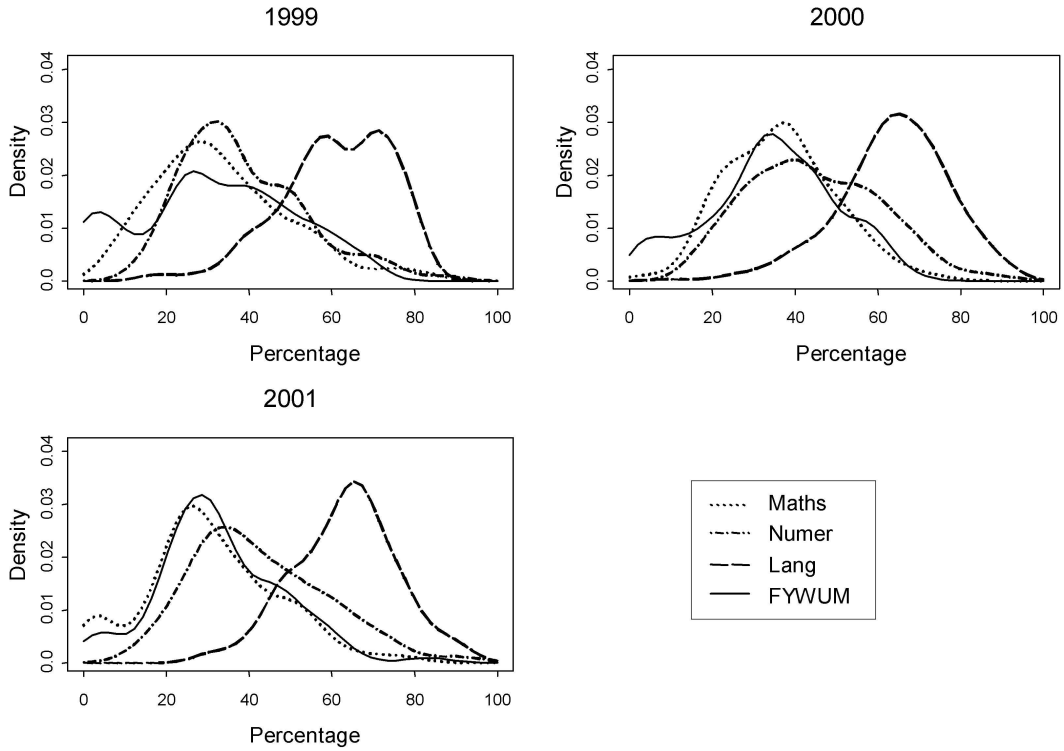

Figure 4: *Kernel density estimates of the densities associated with the access test variables and the FYWUM of the Test Battery 2 data sets for 1999, 2000 and 2001.*

(or densities) with a higher mean. It could then be conjectured from the distributions of FYWUM for 2000 and 2001 that the students are sampled from a normal population with a mean of about 30%, while the rest of the students come from a normal population with a mean of about 50% (2001) or 60% (2000). There seem to be two definite groups of students: a larger unsuccessful group of students that mainly have a FYWUM of less than 50%, and a smaller group of whom at least half have a FYWUM of more than 50%. It should be kept in mind, however, that the effectiveness and accuracy of the assumption of normality needs to be inspected.

Figure 5 depicts alarming results: even though almost all students have Grade 12 and Grade 11 averages of more than 55%, only a few of them achieved a FYWUM of 50% or more. This clearly indicates that school results are creating unrealistically high expectations of university performance. The Grade 11 and 12 averages have their main peaks at approximately 65% over the three years and show very little variation compared to FYWUM.

Although it is standard statistical practice to approximate the distribution of a variable by a normal distribution for calculating probabilities, such a practice could result in inaccurate estimates when the underlying distribution is skewed and/or bimodal. Since the kernel density provides a non-parametric estimate of the distribution of a variable, this suggests using the kernel density for calculating probabilities. Specified cumulative probabilities of the access test variables, Grade 11 (Ave.11) and Grade 12 (Ave.12) final mark and FYWUM were computed using a numerical integration algorithm. Tables 1 and 2

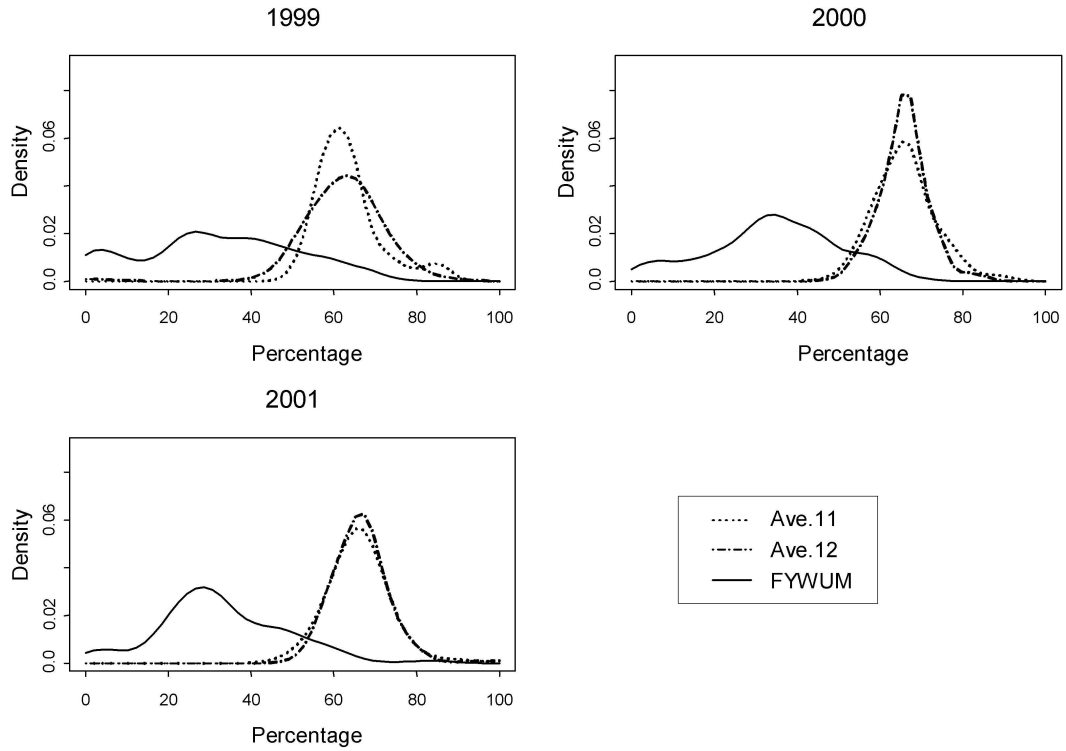

Figure 5: *Kernel density estimates of the densities associated with the Grade 11 and Grade 12 mark variables and the FYWUM of the Test Battery 2 data sets for 1999, 2000 and 2001.*

allow a comparison of probabilities computed by numerical integration of a kernel density estimate and probabilities computed by means of a normal estimate. These tables may be interpreted as follows: the first entry in the Maths column, for example, represents the probability of obtaining a Maths score of at most 10%, the second entry is the probability of obtaining a Maths score of at most 20%, *etc.*

Prob	Maths	Numer	Lang	Ave.11	Ave.12	FYWUM
10%	0.0553	0.0061	0.0005	0.0001	0.0120	0.1704
20%	0.2152	0.0809	0.0094	0.0001	0.0125	0.2722
30%	0.4622	0.3210	0.0236	0.0001	0.0126	0.4635
40%	0.6861	0.5969	0.0768	0.0001	0.0159	0.6474
50%	0.8309	0.7830	0.2010	0.0122	0.0845	0.8060
60%	0.9234	0.9006	0.4421	0.3672	0.3817	0.8679
70%	0.9606	0.9534	0.7020	0.8422	0.7901	0.9342
80%	0.9840	0.9859	0.9439	0.9388	0.9607	0.9497
90%	0.9960	0.9968	0.9995	0.9971	0.9910	0.9500

Table 1: *Test Battery 2 variables of 1999 intake group: Cumulative probabilities calculated by numerical integration.*

In order to verify visually the differences in corresponding probabilities of Tables 1 and 2, Figure 6 was constructed. The normal density estimate, together with the corresponding kernel density estimate, was constructed for each of the variables Maths, Numer, Lang, FYWUM, Ave.11 and Ave.12 for the Test Battery 2 data set of 1999.

Prob	Maths	Numer	Lang	Ave.11	Ave.12	FYWUM
10%	0.0734	0.0263	0.0001	0.0000	0.0000	0.1248
20%	0.2008	0.1032	0.0010	0.0000	0.0000	0.2662
30%	0.4105	0.2780	0.0095	0.0000	0.0006	0.4613
40%	0.6504	0.5343	0.0561	0.0013	0.0126	0.6663
50%	0.8410	0.7766	0.2025	0.0417	0.1105	0.8307
60%	0.9464	0.9245	0.4693	0.3296	0.4165	0.9311
70%	0.9869	0.9826	0.7513	0.8021	0.7889	0.9778
80%	0.9977	0.9973	0.9243	0.9838	0.9653	0.9944
90%	0.9997	0.9997	0.9857	0.9997	0.9977	0.9989

Table 2: *Test Battery 2 variables of 1999 intake group: Cumulative probabilities derived from normal distributions.*

Figure 6 gives an indication of how well the normal density estimation fits the underlying data. It is seen that the normal densities estimate the underlying distribution of Ave.12 rather accurately, but the approximation is less accurate for the variables Maths, Numer, Lang, FYWUM and Ave.11. The kernel estimates indicate that the latter five variables have skewed distributions, which make normal approximation inadequate. Furthermore, it is apparent that the underlying distributions of these five variables have more than one mode. As discussed earlier, these densities seem to be mixtures of two or more, possibly normal, densities.

Since the kernel density estimate can be regarded as a more accurate representation of the true density of a variable, it is clear from Figure 6 that normal probabilities could easily under- or overestimate true probabilities. Investigating, for example, the probability that a student obtains a mark of at most to 30% for Maths, reveals a difference of 0.46 – 0.41 = 0.05 between the kernel and normal density estimates. This is confirmed by Figure 6. Furthermore, if the probability of scoring at most 40% on Numer needs to be ascertained, the normal density estimate provides an answer of 0.53, while the kernel density estimate gives 0.60. Considering the density estimates of Lang, it is confirmed by Tables 1 and 2 that the probability of obtaining a mark of more than 70% is 0.25 for normal estimation while it is 0.30 for kernel estimation.

The FYWUM graph in Figure 6 once again stresses the importance of kernel density estimation. Consider the left tail: the normal density in comparison to the kernel density estimate underestimates the probability that a student would receive a first-year mark of at most 10% by 0.17 – 0.12 = 0.05. The difference between the estimated kernel and normal probabilities of obtaining a FYWUM in excess of 60% is 0.13 – 0.07 = 0.06, indicating that the estimate based on the kernel density is almost twice as large as that based on the estimated normal density. Inspection of Ave.11 reveals that the normal density in comparison to the kernel density estimate overestimates the probability that a student would receive a mark higher than 60% or 70% by 0.67 – 0.63 = 0.04 and 0.08 – 0.04 = 0.04, respectively.

Apart from comparing kernel and normal density estimates, the probabilities calculated by numerical integration may also be used to measure quantitatively to what extent the density estimates of the access test variables correspond to the density estimates of FYWUM. Indeed, Table 1 confirms that the cumulative probabilities of Maths match the correspond-

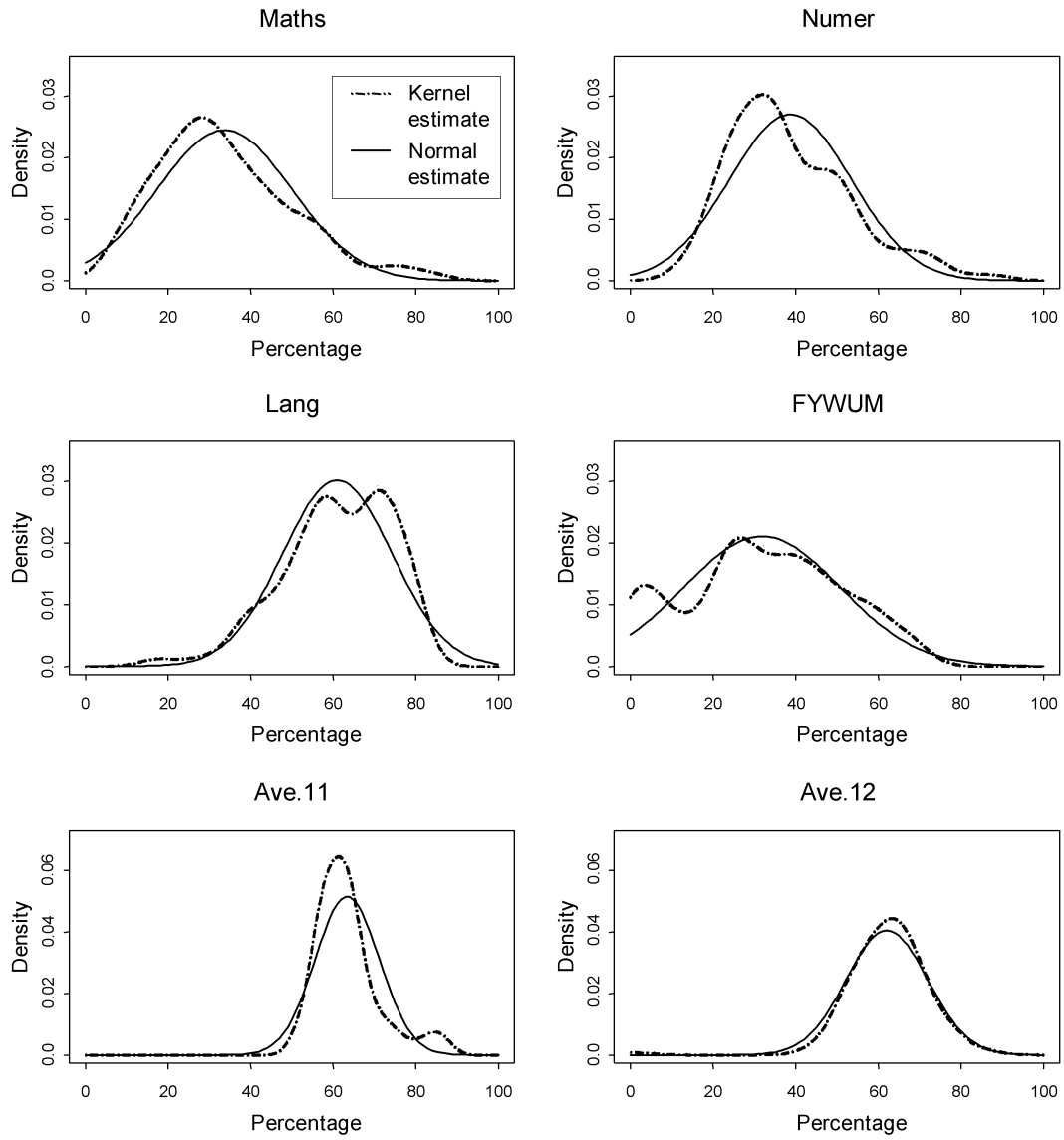

Figure 6: *Parametric normal and non-parametric kernel density estimates for selected variables of Test Battery 2, 1999 intake group.*

ing probabilities of FYWUM reasonably closely. A similar comparison shows that the differences between FYWUM and the school result variables are much more pronounced.

2.3 Density estimates of variables associated with Test Battery 3

Finally, the densities of variables corresponding to the access test, school and first-year performance marks of students who wrote Test Battery 3 are estimated and displayed in Figures 7 and 8.

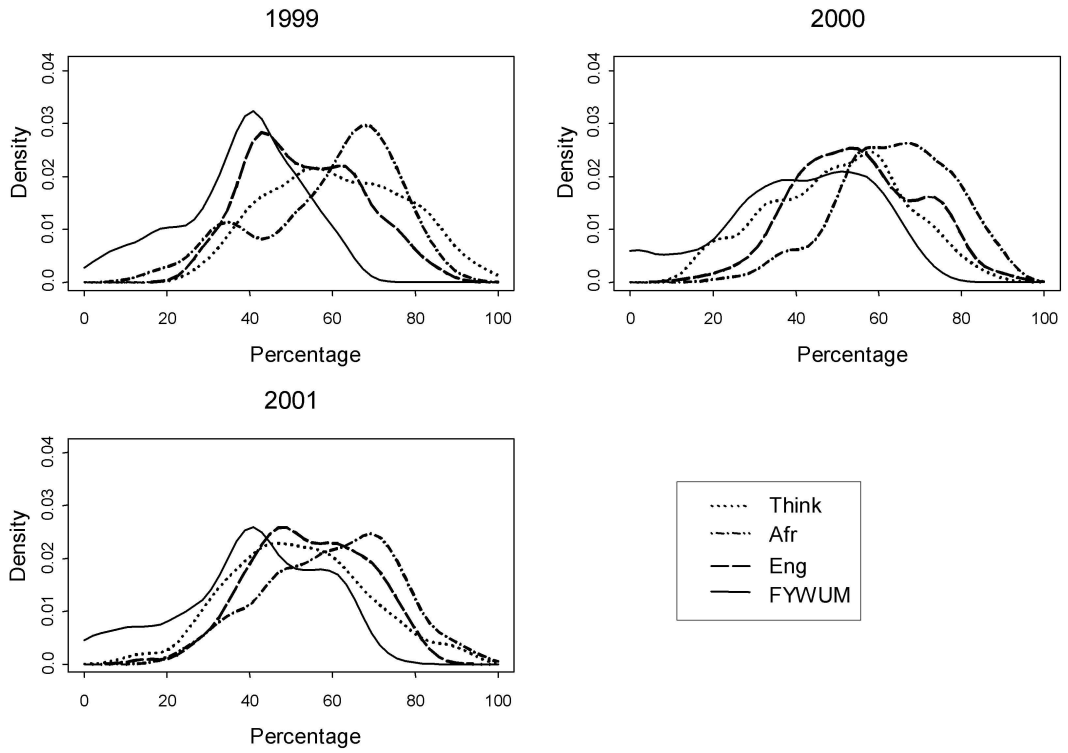

Figure 7: *Kernel densities of the access test variables and the FYWUM of the Test Battery 3 data sets for 1999, 2000 and 2001.*

The mean of FYWUM is calculated as 40% in both 2000 and 2001. The kernel density estimates of FYWUM for 2000 and 2001, however, suggest that the students in these data sets are sampled from a mixture of two populations: either one with a mean of about 40% or another population with a mean of approximately 60%. This stresses again the danger of using normal density estimates in these cases.

Variable Afr is clearly skewed to the left, with a mode at about 70%. Calculation of cumulative probabilities by numerical integration reveals that more than half of the students also obtained a Think score in excess of 50%, and the same holds for the Eng scores. The main peak of FYWUM is about 40% in 1999 and 2001, but about 50% in 2000, although it is clear that more than half of the students have a FYWUM mark of less than 50%. It is seen, therefore, that the distributions of the access test variables do not correspond so closely to that of FYWUM, as was the case in Test Batteries 1 and 2.

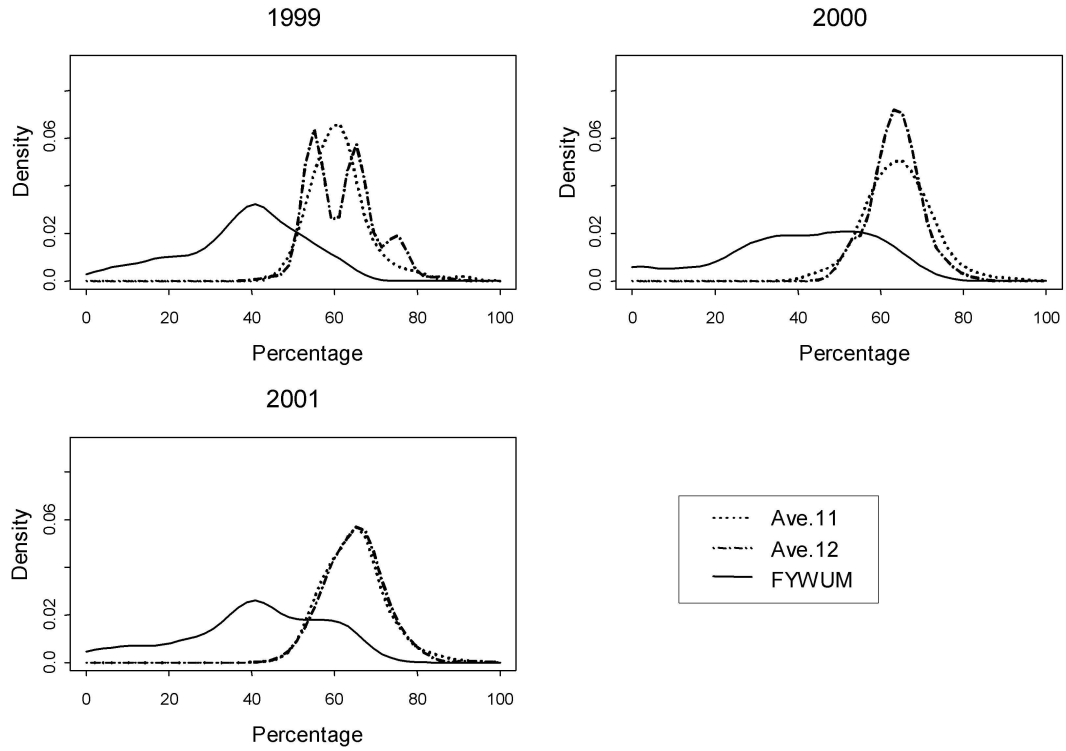

Figure 8: *Kernel densities of the Grade 11 and Grade 12 mark variables and the FYWUM of the Test Battery 3 data sets for 1999, 2000 and 2001.*

It is seen from Figure 8 that in 2000 and 2001 the peaks of Ave.11 and Ave.12 are approximately 65%, while in 1999 Ave.11 has a peak at about 60%, with Ave.12 having two main peaks, one at about 55% and the other at about 65%. As with the results of Test Batteries 1 and 2, although nearly all the Test Battery 3 students achieved Ave.11 and Ave.12 scores of more than 50%, only a small percentage of these students obtained a FYWUM of 50% or more.

3 The bagplot: a bivariate boxplot

Several attempts have been made to generalise the univariate boxplot to two dimensions [3, 4, 12]. In this investigation the bagplot, proposed by Rousseeuw *et al.* (1999a), is introduced for analysing several bivariate distributions involving FYWUM. A bagplot may be regarded as a bivariate boxplot. The main components of a bagplot are a bag that contains the inner 50% of the data points, a fence that separates inliers from outliers, and a loop indicating the points outside the bag, but inside the fence. The bivariate location of the data is expressed in terms of the depth median [2, 9] and will be indicated by a cross. Note that the bag (represented in the graph by a polygon with dark grey interior) corresponds to the box of the univariate case, both containing the 50% of data points closest to the depth median or median, respectively. The area outside the bag but inside the fence (indicated by a light grey loop) plays the same role as the whiskers in one dimension, and the area beyond this loop contains the (bivariate) outliers (indicated by

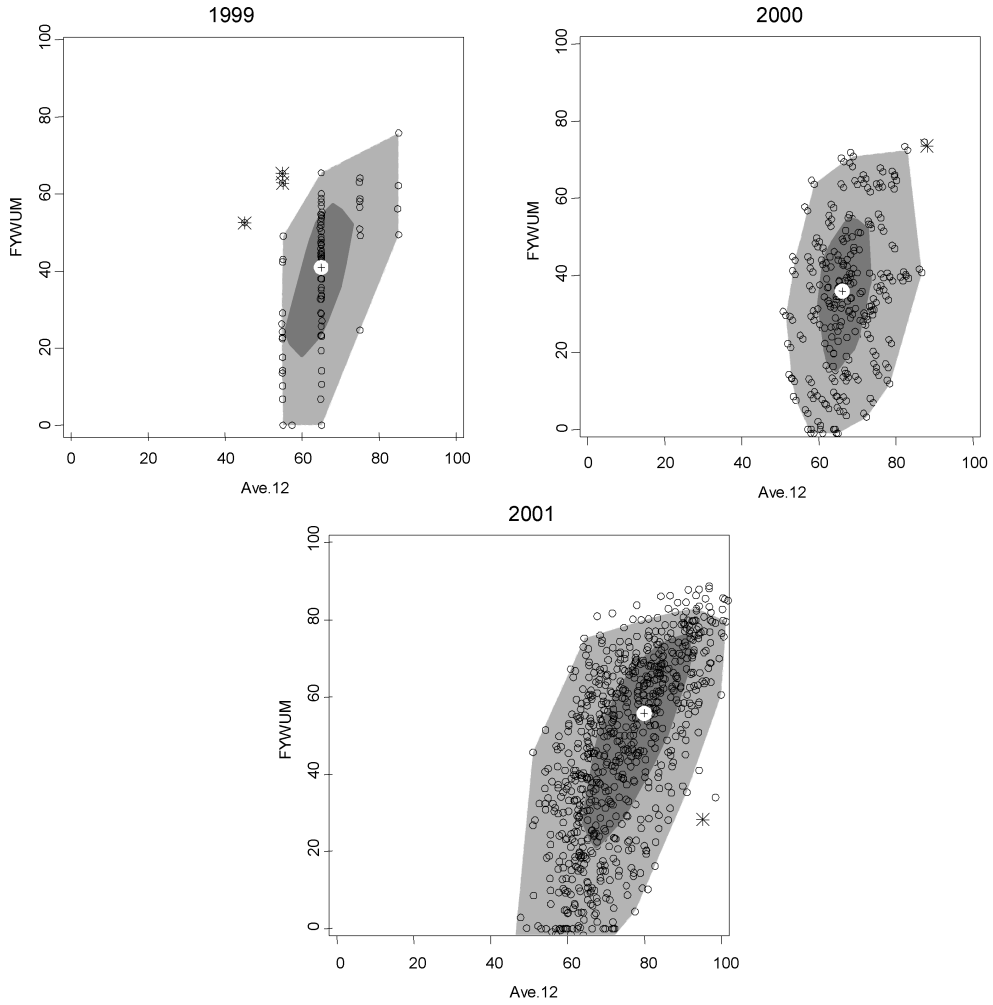

Figure 9: *Bagplots of FYWUM and Ave.12 for 1999, 2000 and 2001.*

black stars). The bagplot thus gives a visual summary of the location (the depth median), spread (the size of the bag), correlation (the orientation of the bag), skewness (the shape of the bag and the loop), and tails (the points near the boundary of the loop and the outliers) of a bivariate data set.

3.1 Constructing bagplots for access test data sets

An S-Plus function for constructing a bagplot is provided by Rousseeuw *et al.* (1999b). This function, however, cannot construct a bagplot if too many points coincide [11]. The three Test Battery data sets typically have many data points that coincide. Therefore the S-Plus function provided by Rousseeuw *et al.* (1999b) was adjusted by adding a small random normal value to each variable concerned. Since the computation time for calculating the depth median increases exponentially with the sample size n, an approximation is used when n is larger than 150 [5]. In this case, the depth median and bag are computed for a random subsample of size 150. The other computations, whose times are linear in

the sample size, are performed on the full data set.

Bagplots were constructed to examine the statistical properties of selected bivariate distributions of the data sets graphically. In this article bagplots for only the Test Battery 1 data set are reported, *viz.* FYWUM and Ave.12 as well as FYWUM and Science (for the 1999, 2000 and 2001 intake groups). These graphs are given in Figures 9 and 10, respectively. For comparison purposes the Tukey medians of the three bivariate distributions displayed in Figure 9 are reproduced in Table 3 together with the corresponding means.

Table 3 reveals that the bivariate means and depth medians correspond well in 1999 and 2000, but differ to some extent in 2001. Note that all the depth medians and bivariate means have a higher Ave.12 coordinate than FYWUM coordinate. In fact, the differences between these two coordinates are always more than 20 percentage points. In 2001 the average FYWUM is larger than 50% for the first time, and the Grade 12 average mark is also reasonably higher than before (larger than 75%). This indicates that the change in the composition of the 2001 intake group (the *effect of the Health Sciences 2001 students*) has an influence on the bivariate distribution of Ave.12 and FYWUM.

Intake	Means		Tukey medians	
group	Ave.12	FYWUM	Ave.12	FYWUM
1999	64.66	39.39	64.99	42.58
2000	66.78	34.89	66.18	36.05
2001	75.77	50.03	79.23	56.04

Table 3: *Means and Tukey medians of Ave.12 and FYWUM for 1999, 2000 and 2001.*

The size of the bag and the loop in Figure 9 shows a substantial increase from 2000 to 2001, indicating that this two-dimensional distribution has markedly larger variation in 2001 than in 2000. The 1999 and 2001 bags show an upward slope, while the 2000 bag is orientated almost vertically. These slopes reflect the sample correlations between Ave.12 and FYWUM calculated as 0.45, 0.34 and 0.67 in 1999, 2000 and 2001 respectively. The bivariate distributions of 1999 and 2001 are slightly skewed, since the 1999 bag is positioned towards the left side and the 2001 bag towards the upper side of the corresponding larger loops. This indicates that the 1999 data set contains only a few students with scores high in both Ave.12 and FYWUM, but the 2001 data set has a relatively large number of students with high scores on both variables.

The bagplots in Figure 9 reveal a weak tendency among first-year students with higher Grade 12 marks to perform better at university. However, the *effect of the Health Sciences 2001 students* indicates that this tendency improves notably if more students with high Grade 12 averages (more than 70%) are required to write access tests.

Table 4 contains the coordinates of the means and depth medians associated with the graphs in Figure 10.

Note that in 2001 both coordinates of the Tukey medians exceed those of the means. From Table 4 the increase in the Science coordinates over the three years is evident, while the FYWUM coordinates show a substantial increase from 2000 to 2001. Comparing Table 4 with Table 3, it is clear that the Science coordinate and FYWUM coordinate, for each of the means and Tukey medians, do not differ to the same extent as the Ave.12 and FYWUM

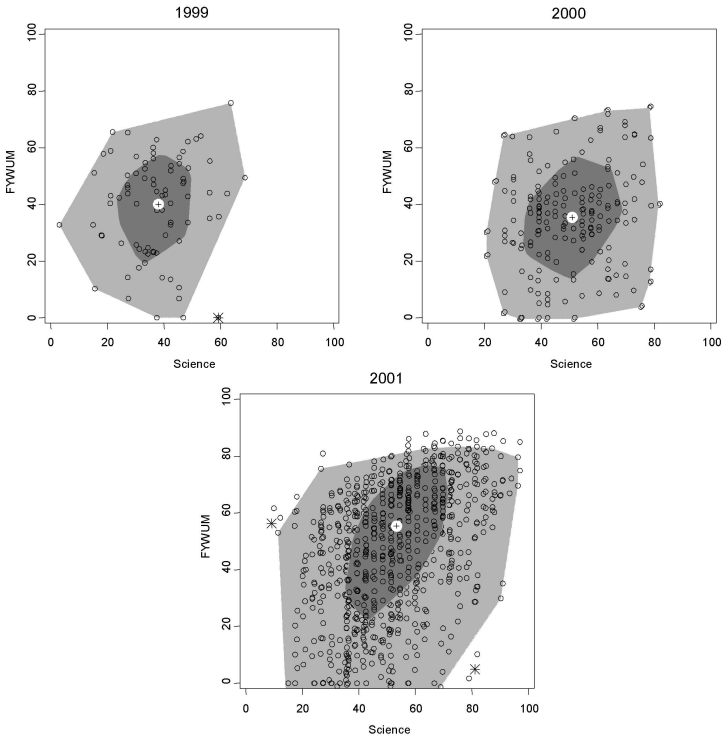

Figure 10: *Bagplots of FYWUM and Science for 1999, 2000 and 2001.*

coordinates. This is in agreement with Figure 2, which reveals a close correspondence in the univariate density estimates of FYWUM and Science. Furthermore, since the corresponding means and Tukey medians are nearly identical for the 1999 intake group as well as the 2000 intake group, it follows that both the 1999 and 2000 bivariate distributions of Science and FYWUM are less skewed than the corresponding 2001 distribution.

A striking feature of the 2001 bivariate distribution (Figure 10) is the upward slope of the bag and long tails in the direction of low scores on both variables, contrary to the more circular bags of 1999 and 2000. It is evident, therefore, that the correlation between FYWUM and Science is the highest for the 2001 intake group. This agrees with the correlations between FYWUM and Science calculated as 0.10, 0.20 and 0.44 for the 1999, 2000 and 2001 intake groups, respectively. Note the increase in the size of the loop for both bivariate distributions in 2001, indicating an increase in variation in the bivariate distribution.

Intake	Means		Tukey medians	
group	Ave.12	FYWUM	Ave.12	FYWUM
1999	37.39	39.39	38.31	39.97
2000	50.64	34.89	50.96	35.51
2001	52.64	50.03	53.49	55.60

Table 4: *Means and Tukey medians of Science and FYWUM for 1999, 2000 and 2001.*

4 Conclusion

Initial analyses were conducted in order to study the univariate properties of the access test, school result and FYWUM variables, and also to compare the properties of these variables

The underlying densities of the access test variables, the school result variables and FYWUM were estimated. The article illustrated that a significant amount of vital information, such as the skewness and modality, about the true underlying distribution may be lost by the common practice of fitting parametric normal density estimates to variables. Furthermore, the estimation of probabilities based on normal densities could therefore result in the underestimation or overestimation of the true probabilities. Indeed, this was illustrated for the Test Battery 2 1999 intake group. This problem is addressed by fitting non-parametric kernel density estimates to the respective distributions, which not only preserve the underlying characteristics of a variable, but also lead to more accurate probability estimation (by means of numerical integration). It is found that the density estimates of Maths, Science and Numer correspond reasonably well with those of FYWUM. The differences in the location and spread of the density estimates of the school result variables, when compared to those of FYWUM, are striking. Contrary to the majority of students obtaining Grade 11 and 12 final marks higher than 50%, only a small percentage of these students received a FYWUM of higher than 50%. This stresses once again that the school results might lead to unrealistic expectations of performance at university. The *effect of the Health Sciences 2001 students* is demonstrated by a pronounced increase in the mean of FYWUM, being in excess of 50% for the first time.

The underlying relationship among the access tests, school results and FYWUM are also taken into consideration. The bagplot provides a graphical display, summarising several important properties of a bivariate distribution. Two bivariate distributions of the Test Battery 1 data set were considered. The upward slope of the bag of the bagplot of FYWUM and Ave.12 suggests a positive correlation between these two variables. The bagplot of FYWUM and Science, however, has a more circular bag, which confirms the lower positive correlation between the two variables concerned. The *effect of the Health Sciences 2001 students* is visually demonstrated by the upward slope and position of the bags of both bivariate distributions, suggesting higher correlations as well as an increase in the number of students who obtained a FYWUM in excess of 50%.

This article clearly demonstrates that the statistical instruments employed to analyse and represent data play a crucial role when variables such as school results, access tests and first-year university performance are explored. Important findings based on these analyses are that the estimated distributions of certain of the access tests (*viz.* Mathematics, Science and Numeracy Skills) correspond reasonably well with those of average first-year university performance. These findings cast doubt upon the ability of average school marks to predict average first-year university performance realistically. Furthermore, it seems that only a very small percentage of those students with a school result average of below 70% obtain a first-year university average performance of 50% or more. The ability of bagplots to provide a visual summary of important features like location, spread, correlation, skewness, outliers and tails of bivariate data sets originating from first-year university average performance and another variable is indeed of value for all decision makers involved in the

admission of students to university, and the schools that prepare the prospective students for university studies.

References

[1] Bothma A, Botha HL & Le Roux NJ, 2004, *School results and access test results as indicators of first-year performance at university*, ORiON, **20(1)**, pp. 73–88.

[2] Donoho DL & Gasko M, 1992, *Breakdown properties of location estimates based on halfspace depth and projected outlyingness*, The Annals of Statistics, **20**, pp. 1803–1827.

[3] Goldberg KM & Iglewicz B, 1992, *Bivariate extensions of the boxplot*, Technometrics, **34**, pp. 307–320.

[4] Liu RY, Parelius JM & Singh K, 1999, *Multivariate analysis by data depth: descriptive statistics, graphics and inference*, The Annals of Statistics, **27**, pp. 783–858.

[5] Rousseeuw PJ, Ruts I & Tukey JW, 1999a, *The bagplot: a bivariate boxplot*, The American Statistician, **53**, pp. 382–387.

[6] Rousseeuw PJ, Ruts I & Tukey JW, 1999b, *Bagplot functions library*, [Online] Available from `http://win-www.uia.ac.be/u/statis/index.html`.

[7] Sanoff AP, 2004, *Some find SATs don't 'define quality'*, USA Today, 01-10-2004. [Online], [Cited: 25 October 2004], Available from `http://www.usatoday.com/news/education/200-10-01-sat_x.htm`

[8] Scott DW, 1992, *Multivariate density estimation: Theory, practice and visualization*, John Wiley and Sons, New York (NY).

[9] Tukey JW, 1975, *Mathematics and the picturing of data*, Proceedings of the International Congress of Mathematics, **2**, pp. 523–531.

[10] Venables WN & Ripley BD, 2002, *Modern applied statistics in S-Plus*, 4^{th} Edition, Springer-Verlag, New York (NY).

[11] Wood LC, 2001, *A statistical analysis of differentiation in the education of South African youth based on the living standards and development survey 1993*, Masters assignment, Stellenbosch University, Stellenbosch.

[12] Zani S, Riani M & Corbellini A, 1998, *Robust bivariate boxplots and multiple outlier detection*, Computational Statistics and Data Analysis, **28**, pp. 257–270.